HELP NGO Assignment

Q1. Assignment Summary

Ans:

Problem Statement: HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities. It runs a lot of operational projects from time to time along with advocacy drives to raise awareness as well as for funding purposes.

After the recent funding program's, they have been able to raise around $ 10 million. Now the CEO of the NGO needs to decide how to use this money strategically and effectively. The significant issues that come while making this decision are mostly related to choosing the countries that are in the direst need of aid.

Objective: Our job is to categorize the countries using some socio-economic and health factors that determine the overall development of the country. Then we need to suggest the countries which the CEO needs to focus on the most.

Technical Approach:

- Using Hierarchical Clustering to identify the optimal cluster value.
- Use Silhouette and Elbow method to validate the optimal cluster value.
- Use K-Means Cluster method to build the final cluster model.
- Identify appropriate cluster for Financial Aid using Cluster Mean method.
- Analyze the Final Cluster Statistics against other clusters.
- Decision making on the final list based on the Descriptive Statistics of our Final Cluster.
- Choose the Top-10 Countries from the Final Cluster based on the Higher Child Mortality, Lower Income and Low GDP.
- Present the Final Report.

Q2. Clustering

Ans:

a) **Compare and contrast K-Means Clustering and Hierarchical Clustering.**

K-means:

- Method of cluster analysis using a pre-specified no. of clusters. It requires advance knowledge of ‘K’.
- K Means clustering needed advance knowledge of K i.e. no. of clusters one want to divide your data.
- One can use median or mean as a cluster center to represent each cluster.
- Methods used are normally less computationally intensive and are suited with very large datasets.

Hierarchical Clustering:

- Also known as hierarchical cluster analysis (HCA) is also a method of cluster analysis which seeks to build a hierarchy of clusters without having fixed number of cluster.
- In hierarchical clustering one can stop at any number of clusters, one find appropriate by interpreting the dendrogram.
- Agglomerative methods begin with ‘n’ clusters and sequentially combine similar clusters until only one cluster is obtained.
- Divisive methods work in the opposite direction, beginning with one cluster that includes all the records and Hierarchical methods are especially useful when the target is to arrange the clusters into a natural hierarchy.

b) **Briefly explain the steps of the K-means clustering algorithm.**

I. Choose the number of cluster k.
   - Pick the desired number of clusters according to the data.

II. Select k random points from the data as centroids.

- We randomly have to select the centroid for each cluster. Let's say we want to have 2 clusters, so k is equal to 2 here. We then have to randomly select the centroid.

III. Assign all the points to the closest cluster centroid.
    - Once we have initialized the centroids, we assign each point to the closest cluster centroid.

IV. Recompute the centroids of newly formed clusters.
    - Once we have assigned all of the points to either cluster, the next step is to compute the centroids of newly formed clusters.

V. Repeat the Steps III and IV.

c) How is the value of 'k' chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.

I. **Elbow Method:** Define clusters such that the total intra-cluster variation [or total within-cluster sum of square (WSS)] is minimized. The total WSS measures the compactness of the clustering and we want it to be as small as possible.
The Elbow method looks at the total WSS as a function of the number of clusters: One should choose a number of clusters so that adding another cluster doesn't improve much better the total WSS.

II. **Silhouette Method:** Determines how well each object lies within its cluster. A high average silhouette width indicates a good clustering.
Average silhouette method computes the average silhouette of observations for different values of k. The optimal number of clusters k is the one that maximize the average silhouette over a range of possible values for k (Kaufman and Rousseeuw 1990).

III. **Gap Statistic:** The gap statistic compares the total within intra-cluster variation for different values of k with their expected values under null reference distribution of the data. The estimate of the optimal clusters will be value that maximize the gap statistic (i.e, that yields the largest gap statistic). This means that the clustering structure is far away from the random uniform distribution of points.

These were some of the points that explains the algorithm approach to find the optimal K value in K-Means Clustering, now explaining the **business aspect** of the K-Means:

The K-means clustering algorithm is used to find groups which have not been explicitly labeled in the data. This can be used to confirm business assumptions about what types of groups exist or to identify unknown groups in complex data sets.

d) Explain the necessity for scaling/standardization before performing Clustering.

In statistics, standardization (sometimes called data normalization or feature scaling) refers to the process of rescaling the values of the variables in your data set so they share a common scale. Often performed as a pre-processing step, particularly for cluster analysis, standardization may be important if you are working with data where each variable has a different unit (e.g., inches, meters, tons and kilograms), or where the scales of each of your variables are very different from one another (e.g., 0-1 vs 0-1000). The reason this importance is particularly high in cluster analysis is because groups are defined based on the distance between points in mathematical space.

Standardization helps to make the relative weight of each variable equal by converting each variable to a unit-less measure or relative distance.

e) Explain the different linkages used in Hierarchical Clustering.

Single-Linkage

Single-linkage (nearest neighbor) is the shortest distance between a pair of observations in two clusters. It can sometimes produce clusters where observations in different clusters are closer together than to observations within their own clusters. These clusters can appear spread-out.

Complete-Linkage

Complete-linkage (farthest neighbor) is where distance is measured between the farthest pair of observations in two clusters. This method usually produces tighter clusters than single-linkage, but these tight clusters can end up very close together. Along with average-linkage, it is one of the more popular distance metrics.

Average-Linkage

Average-linkage is where the distance between each pair of observations in each cluster are added up and divided by the number of pairs to get an average inter-cluster distance. Average-linkage and complete-linkage are the two most popular distance metrics in hierarchical clustering.

Centroid-Linkage

Centroid-linkage is the distance between the centroids of two clusters. As the centroids move with new observations, it is possible that the smaller clusters are more similar to the new larger cluster than to their individual clusters causing an inversion in the dendrogram. This problem doesn’t arise in the other linkage methods because the clusters being merged will always be more similar to themselves than to the new larger cluster.